SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension

Year: 2023 | Citations: 754 | Authors: Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge

Abstract

Based on powerful Large Language Models (LLMs), recent generative Multimodal Large Language Models (MLLMs) have gained prominence as a pivotal research area, exhibiting remarkable capability for both comprehension and generation. In this work, we address the evaluation of generative comprehension in MLLMs as a preliminary step towards a comprehensive assessment of generative models, by introducing a benchmark named SEED-Bench. SEED-Bench consists of 19K multiple choice questions with accurate human annotations (x 6 larger than existing benchmarks), which spans 12 evaluation dimensions including the comprehension of both the image and video modality. We develop an advanced pipeline for generating multiple-choice questions that target specific evaluation dimensions, integrating both automatic filtering and manual verification processes. Multiple-choice questions with groundtruth options derived from human annotation enables an objective and efficient assessment of model performance, eliminating the need for human or GPT intervention during evaluation. We further evaluate the performance of 18 models across all 12 dimensions, covering both the spatial and temporal understanding. By revealing the limitations of existing MLLMs through evaluation results, we aim for SEED-Bench to provide insights for motivating future research. We will launch and consistently maintain a leaderboard to provide a platform for the community to assess and investigate model capability.